

Lumina-T2X: Transforming Text into Any Modality, Resolution, and Duration via Flow-based Large Diffusion Transformers

Year: 2024 | Citations: 118 | Authors: Peng Gao, Le Zhuo, Ziyi Lin, Chris Liu, Junsong Chen

Abstract

Sora unveils the potential of scaling Diffusion Transformer for generating photorealistic images and videos at arbitrary resolutions, aspect ratios, and durations, yet it still lacks sufficient implementation details. In this technical report, we introduce the Lumina-T2X family - a series of Flow-based Large Diffusion Transformers (Flag-DiT) equipped with zero-initialized attention, as a unified framework designed to transform noise into images, videos, multi-view 3D objects, and audio clips conditioned on text instructions. By tokenizing the latent spatial-temporal space and incorporating learnable placeholders such as [nextline] and [nextframe] tokens, Lumina-T2X seamlessly unifies the representations of different modalities across various spatial-temporal resolutions. This unified approach enables training within a single framework for different modalities and allows for flexible generation of multimodal data at any resolution, aspect ratio, and length during inference. Advanced techniques like ROPE, RMSNorm, and flow matching enhance the stability, flexibility, and scalability of Flag-DiT, enabling models of Lumina-T2X to scale up to 7 billion parameters and extend the context window to 128K tokens. This is particularly beneficial for creating ultra-high-definition images with our Lumina-T2I model and long 720p videos with our Lumina-T2V model.

Remarkably, Lumina-T2I, powered by a 5-billion-parameter Flag-DiT, requires only 35% of the training computational costs of a 600-million-parameter naive DiT. Our further comprehensive analysis underscores Lumina-T2X's preliminary capability in resolution extrapolation, high-resolution editing, generating consistent 3D views, and synthesizing videos with seamless transitions. We expect that the open-sourcing of Lumina-T2X will further foster creativity, transparency, and diversity in the generative AI community.